AMENDMENTS TO THE CLAIMS

Upon entry of this amendment, the following listing of claims will replace all prior versions and listings of claims in the pending application.

IN THE CLAIMS

Please amend claims 34, 49 and 64 as follows:

1-33. (Canceled)

34. (Currently Amended) A method for managing throughput while avoiding overload of one or more servers in an environment including an interface unit intercepting requests from clients to a server, transmitting the intercepted requests to the server, and intercepting responses to the requests transmitted by the server to the clients, the method comprising the steps of:

transmitting, by an interface unit, client requests to a server to maintain performance of server throughput within a predetermined threshold range of a server throughput function, the predetermined threshold range corresponding to values of the server throughput function based on a number of client requests to the server via the interface unit per second of time over a total number of clients communicating with the server via the interface unit;

monitoring, by the interface unit, responses to client requests intercepted by the interface unit, changes in response times from the server and changes of a rate in which the response times change;

intercepting, by the interface unit, a request from a client to open a transport layer connection with the server;

determining, by the interface unit, from the monitoring that the performance of the server throughput exceeds the predetermined threshold range;

buffering, by the interface unit in response to the determination, the intercepted request in a queue; and

transmitting, by the interface unit, the buffered request to the server upon the interface unit determining that the performance of server throughput is within the

predetermined threshold range of the server throughput function and that a total number of clients currently communicating with the server per second of time is less than a predetermined number of clients per second of time.

35. (Previously Presented) The method of claim 34, wherein the predetermined threshold range comprises one of a maximum threshold range or an optimal threshold range for server throughput.

36. (Previously Presented) The method of claim 35, wherein the predetermined threshold range comprises a first threshold at a lower point in the predetermined threshold range and a second threshold at a higher point in the predetermined threshold range, the first threshold represents one of a faster response time, a lesser number of users, or a greater number of connections than the second threshold.

37. (Previously Presented) The method of claim 36, comprising transmitting, by the interface unit, client requests to the server to maintain performance of server throughput one of at or near the first threshold.

38. (Previously Presented) The method of claim 34, comprising determining, by the interface unit, the performance of the server throughput based on monitoring one or more of: the number of active connections opened to the server, the response time of the server, the rate at which the response time is changing, and the intercepted request.

39. (Previously Presented) The method of claim 34, determining, by the interface unit, the performance of the server throughput based on a first portion of server resources available to service existing clients and a second portion of server resources available to accept new clients.

40. (Previously Presented) The method of claim 34, comprising identifying a preferred client value for the request of the client, and determining the position of the client request in the queue based on the preferred client value.

41. (Previously Presented) The method of claim 40, comprising determining, by the interface unit, the preferred client value from one or more of the internet address of the client request, the port number of the client request, by a header related to the client request, by previous requests from the client of the client request, and by a cookie related to the client request.

42. (Previously Presented) The method of claim 34, comprising pooling, by the interface unit, a plurality of transport layer connections to the server.

43. (Previously Presented) The method of claim 42, comprising multiplexing, by the interface unit, client requests via the pooled plurality of transport layer connections.

44. (Previously Presented) The method of claim 34, comprising closing, by the interface unit, transport layer connections to the server to bring performance of server throughput within the predetermined threshold range.

45. (Previously Presented) The method of claim 34, comprising determining, by the interface unit, the performance of server throughput by one of a number of requests pending at the server or server error/overload messages from the server.

46. (Previously Presented) The method of claim 34, comprising establishing, by the interface unit, the transport layer connection with the client in response to the request from the client.

47. (Previously Presented) The method of claim 34, comprising opening, by the interface unit, a second transport layer connection to the server if there is not a free transport layer connection to the server.

48. (Previously Presented) The method of claim 34, wherein monitoring further comprises the interface unit monitoring changes in times between the interface unit

forwarding intercepted client requests to the server and the interface unit receiving the responses to the forwarded intercepted client requests from the server and changes of a rate in which differences between the changes in times change.

49. (Currently Amended) A system having an interface unit for managing throughput while avoiding overload of one or more servers, the interface unit intercepting requests from clients to a server, transmitting the intercepted requests to the server, and intercepting responses to the requests transmitted by the server to the clients, the system comprising:

a device comprising a processor and configured as an interface unit ~~for~~ transmitting client requests to a server to maintain performance of server throughput within a predetermined threshold range of a server throughput function, the predetermined threshold range corresponding to a value of the server throughput function based on a number of client requests to the server via the interface unit per second of time over a total number of clients communicating with the server via the interface unit, and intercepting a request from a client to open a transport layer connection with the server; and

a queue for storing intercepted client requests;

wherein the interface unit determines from monitoring responses to client requests intercepted by the interface unit, changes in response times of the server and changes of a rate in which the response times change, that the performance of the server throughput exceeds the predetermined threshold range, and buffers in the queue the intercepted request in response to the determination; and

wherein the interface unit transmits the buffered request upon determining that the performance of server throughput is within the predetermined threshold range of the server throughput function and that a total number of clients currently communicating with the server per second of time is less than a predetermined number of clients per second of time.

50. (Previously Presented) The system of claim 49, wherein the predetermined threshold

range comprises one of a maximum threshold range or an optimal threshold range for server throughput.

51. (Previously Presented) The system of claim 50, wherein the predetermined threshold range comprises a first threshold at a lower point in the predetermined threshold range and a second threshold at a higher point in the predetermined threshold range, the first threshold represents one of a faster response time, a lesser number of users, or a greater number of connections than the second threshold.

52. (Previously Presented) The system of claim 51, wherein the interface unit transmits client requests to the server to maintain performance of server throughput one of at or near the first threshold.

53. (Previously Presented) The system of claim 49, wherein the interface unit determines the performance of the server throughput based on monitoring one or more of: the number of active connections opened to the server, the response time of the server, the rate at which the response time is changing, and the buffered request.

54. (Previously Presented) The system of claim 49, wherein the interface unit determines the performance of the server throughput based on a first portion of server resources available to service existing clients and a second portion of server resources available to accept new clients.

55. (Previously Presented) The system of claim 49, wherein the interface unit identifies a preferred client value for the request of the client, and determines the position of the client request in the queue based on the preferred client value.

56. (Previously Presented) The system of claim 49 ,wherein the interface unit determines the preferred client value from one or more of the internet address of the client request, the port number of the client request, by a header related to the client request, by previous requests from the client of the client request, and by a cookie related to the client request.

57. (Previously Presented) The system of claim 49, wherein the interface unit pools a plurality of transport layer connections to the server.

58. (Previously Presented) The system of claim 57, wherein the interface unit multiplexes client requests via the pooled plurality of transport layer connections.

59. (Previously Presented) The system of claim 49, wherein the interface unit closes the transport layer connection to the server to bring performance of server throughput within the predetermined threshold range.

60. (Previously Presented) The system of claim 49, wherein the interface unit determines the performance of server throughput by one of a number of requests pending at the server or server error/overload messages from the server.

61. (Previously Presented) The system of claim 49, wherein the interface unit establishes the transport layer connection with the client in response to the request from the client.

62. (Previously Presented) The system of claim 49, wherein the interface unit opens a second transport layer connection to the server if there is not a free transport layer connection to the server.

63. (Previously Presented) The system of claim 49, wherein the interface unit opens a second transport layer connection to the server if the queue comprises one or more requests from a second client.

64. (Allowed) A method for managing throughput while avoiding overload of one or more servers in an environment including an interface unit intercepting requests from clients to a server, transmitting the intercepted requests to the server, and intercepting responses to the requests transmitted by the server to the clients, the method comprising the steps of:

transmitting, by a device configured as an interface unit, client requests to a server to maintain performance of server throughput within a predetermined threshold of a server throughput function, the predetermined threshold corresponding to a value of the server throughput function that is based on a number of client requests to the server via the interface unit per second of time over a total number of clients communicating with the server via the interface unit;

monitoring, by the interface unit, changes in times between the interface unit forwarding intercepted client requests to the server and the interface unit receiving the responses to the forwarded intercepted client requests from the server and changes of a rate in which differences between the changes in times change;

intercepting, by the interface unit, a request from a client to open a connection with the server;

determining, by the interface unit, from the monitoring that the performance of the server throughput exceeds the predetermined threshold ~~range~~;

buffering, by the interface unit in response to the determination, the intercepted request in a queue; and

transmitting, by the interface unit, the buffered request to the server upon the interface unit determining that the performance of server throughput is within the predetermined threshold of the server throughput function and that a total number of clients currently communicating with the server per second of time is less than a predetermined number of clients per ~~the per~~ second of time.